



STScI | SPACE TELESCOPE
SCIENCE INSTITUTE

EXPANDING THE FRONTIERS OF SPACE ASTRONOMY

JupyterHub on AWS

Tom Donaldson, on behalf of
Christian Mesh, Mike Fox, Iva Momcheva,
Josh Peek, Arfon Smith

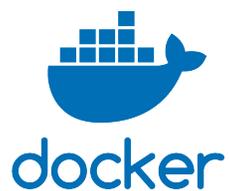


Highlights of what we're building

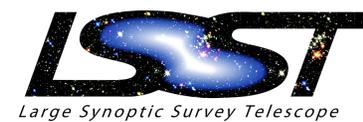
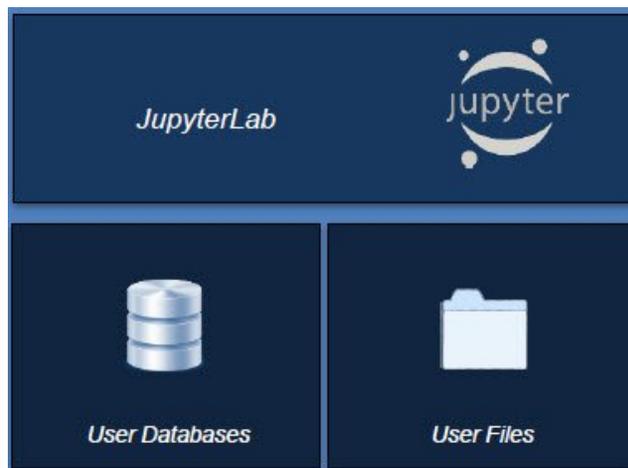
- Cloud-hosted copy of all HST public data
- (Live) JupyterLab environment with *some* compute/storage
- Collection of Docker containers installed with common tools



Common technologies, many implementers



kubernetes





Amazon Web Services: Public Dataset Program

Large Datasets Repository | Pl: x Arlon

Secure | <https://aws.amazon.com/public-datasets/>

Menu **aws** Contact Sales Products Solutions Pricing More English My Account Sign In to the Console

RELATED LINKS

- Big Data on AWS
- Open Data on AWS
- AWS Programs for Research and Education

AWS Public Datasets

Sign up now

AWS hosts a variety of public datasets that anyone can access for free.

Previously, large datasets such as satellite imagery or genomic data have required hours or days to locate, download, customize, and analyze. When data is made publicly available on AWS, anyone can analyze any volume of data without needing to download or store it themselves. These datasets can be analyzed using AWS compute and data analytics products, including [Amazon EC2](#), [Amazon Athena](#), [AWS Lambda](#) and [Amazon EMR](#).

Available Public Datasets on AWS

Geospatial and Environmental Datasets

Learn more about working with geospatial data on AWS at [Earth on AWS](#).

- Landsat on AWS:** An ongoing collection of satellite imagery of all land on Earth produced by the Landsat 8 satellite.
- Sentinel-2 on AWS:** An ongoing collection of satellite imagery of all land on Earth produced by the Sentinel-2 satellite.
- GOES on AWS:** GOES provides continuous weather imagery and monitoring of meteorological and space environment data across North America.
- SpaceNet on AWS:** A corpus of commercial satellite imagery and labeled training data to foster innovation in the development of computer vision algorithms.
- OpenStreetMap on AWS:** OSM is a free, editable map of the world, created and maintained by volunteers. Regular OSM data archives are made available in Amazon S3.
- MODIS on AWS:** Select products from the Moderate Resolution Imaging Spectroradiometer (MODIS) managed by the U.S. Geological Survey and NASA.

- ~120TB public HST data for ACS, COS, STIS, WFC3, FGS
- Range of high-impact datasets
- Hosted in cloud - 'highly available'
- Enable new types of data analyses
- Hosted at no cost to STScI/NASA



Amazon Web Services: Public Dataset Program

- Data is hosted in an S3 region (for ‘free’)
- Conditions of inclusion in program: make the data useful:
 - An AMI with a demonstration of how to use the public dataset must be provided
 - AWS recover costs by making access to the data free from AWS services (EC2), making it cost effective for researchers to buy AWS computing time
 - Enables new types of analyses



JupyterLab environment

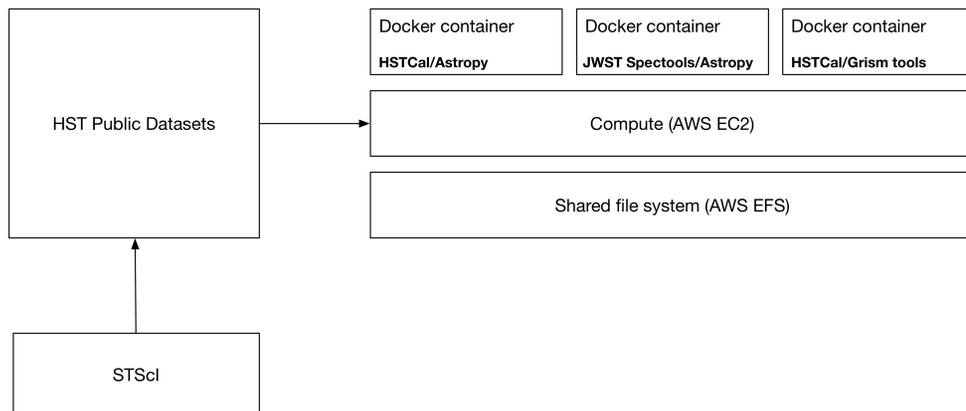
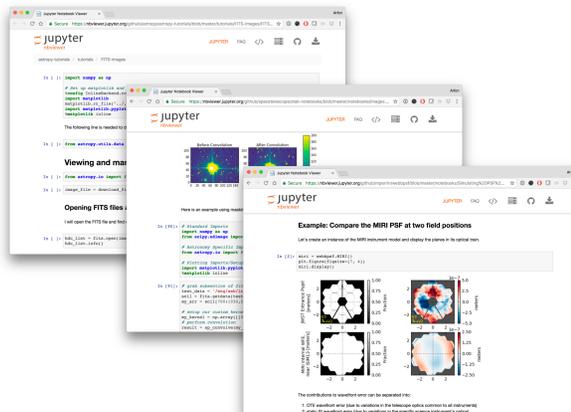
The screenshot displays the JupyterLab interface. On the left, there is a sidebar with navigation options like 'CONSOLE', 'EDITOR', and 'FILE OPERATIONS'. The main area is divided into three panes:

- Code Editor:** Contains Python code for data analysis and plotting. The code includes imports for `matplotlib`, `numpy`, and `matplotlib.pyplot`. It defines a function `plot_beta_hist` and uses it to generate histograms. It also includes code for loading and plotting MRI data.
- Console:** Shows the output of the code execution, including the text: `In [1]:` followed by the function definition and the output of the `plot_beta_hist` function calls.
- Figure:** Displays a histogram with three overlapping distributions in green, red, and purple. Below the histogram, there is a small image of a brain and a line plot showing MRI intensity over time.

- Interactive computing environment
- Where most development work is going from the core Jupyter team
- Works with community tools (e.g. Astropy)



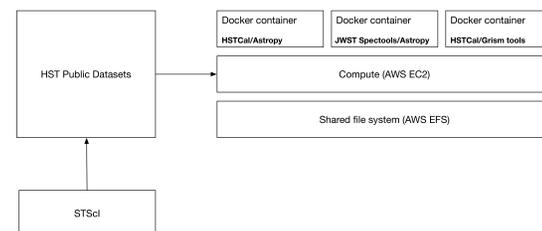
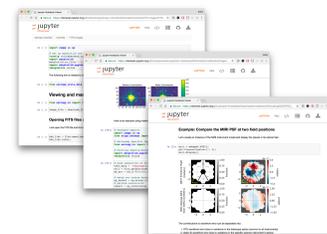
Containers for different environments





Technical details:

- JupyterHub, a multi-user Hub, spawns, manages, and proxies multiple instances of the single-user Jupyter notebook server.
- JupyterLab frontend provides notebook server, file management, and a terminal shell
- Using Docker to containerize science environments, allows a verified computing environment to be instantiated rapidly.
- Containers are versioned providing precise reproducibility of the research environment
- AWS computing resources scale with user load, providing good cost efficiency
- Container orchestration provides high availability, healing the cluster when there are hardware failures





Core technical challenges

- Creating containers with pipeline/common software stacks
- Managing the cloud environment well:
 - User quotas (storage, compute etc.)
 - User storage (home directories), backups
 - Scalable, highly-available infrastructure (with cost caps/alerts)
- Relative inexperience of STScI with commercial cloud

** Relatively few large-scale JupyterHub deployments on AWS



Jupyter on AWS

- Started with Zero to JupyterHub
 - <http://zero-to-jupyterhub.readthedocs.io/en/latest/>
 - For installing and managing JupyterHub with Kubernetes
 - Documentation was sparse.
 - Many teething pains, especially in managing storage.
 - There were hundreds of lines of comments in k8s regarding how to handle storage, but with no decisions.
 - Worked through many issues, then contributed back rewritten documentation and code (via open source pull requests).
- Goal is to have an out-of-the-box solution anyone could spin up.
- Users have begun exercising the system.
- Automated tests being added to really push the system, especially with scalability.
- Using github authN now, but work has begun on integrating with STScI SSO
 - STScI SSO uses Shib, but we will probably create a bridge to that to leverage predominance of OAuth elsewhere.